Behavior Genetics, Vol. 35, No. 4, July 2005 (© 2005)
DOI: 10.1007/s10519-005-2791-5

Mx Scripts Library: Structural Equation Modeling Scripts for Twin and Family Data

D Posthuma[1] and D. I. Boomsma[1]

Received 16 Aug. 2004—Final Jan. 10 2005

Structural equation modeling (SEM) provides a flexible tool to carry out genetic analyses of family and twin data. The basic model which decomposes the variance between and within families for a particular trait into genetic and non-genetic components can be generalized to multivariate and/ or longitudinal data, incorporate sex differences in parameter estimates, and model the effects of measured environment, candidate genes or DNA marker data. We introduce a web-based library (http://www.psy.vu.nl/mxbib) of scripts for uni- and multivariate genetic epidemiological analyses, as well as for linkage and genetic association tests. The scripts are written to be used with the freely available software package Mx and provide a flexible and uniform approach to the analysis of data from relatives.

KEY WORDS: Genetic analysis; Mx; structural equation modeling; twin and family data.

INTRODUCTION

The pattern of correlations between relatives for uni- or multivariate traits provides information on the underlying sources of variation and covariation in those traits (Boomsma *et al*., 2002). For example, in a sample of monozygotic (MZ) and dizygotic (DZ) twins, twice the difference between the MZ and DZ twin correlations for a particular trait provides a first estimate of the heritability of that trait. Likewise, in a sample of unrelated subjects who are raised together, i.e., adoptees, the importance of the shared family environment is given by the correlation between genetically unrelated siblings. The proportional contribution of the non-shared environmental influences to variation in a certain trait is given by the extent to which MZ twins do not resemble each other.

For quantitative traits, the product-moment or alternatively the intra-class correlation of trait values is obtained from the raw, or transformed, data. For categorical traits (e.g., dichotomies as affected/ unaffected) correlations between relatives on the underlying scale of liability (i.e., tetrachoric and polychoric correlations) are often employed (Falconer, 1989). However, these approaches are descriptive and do not provide information on how well the calculated parameters describe the observed data, nor do they allow actual testing of different hypotheses (Eaves, 1969). In addition, for traits for which the heritabilities differ between sexes, it is hard to use data from opposite-sex relatives. Moreover, the correlational approach is difficult to generalize to multivariate data, in which the researcher may wish to explore alternative multivariate models to explain the pattern of associations between traits.

The analysis of covariance structures by using iterative procedures, such as maximum likelihood estimation, allows evaluation of how well a theoretical model describes the observed data (Eaves, 1969; Jinks and Fulker, 1970). The analysis of covariance structures is the main focus of SEM. SEM has several advantages over merely comparing the MZ and DZ correlations; SEM allows parameter *estimation*, while the correlational method merely allows parameter *calculation*. SEM thus also allows determination of confidence intervals and of standard errors of parameter estimates and quantifies how well a model describes the data. SEM also allows testing models overall rather than coefficients individually, as well as

[1] Department of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands.

specification of (non)-scalar sex-limitation models, specification of (non)-scalar age-limitation models, and the inclusion of fixed effects of subpopulations, e.g., according to SES or medication status. In addition, SEM can incorporate models which include social interactions between family members, which predict differences in variances between, e.g., MZ and DZ twins (Eaves, 1976).

SEM has been implemented in a number of widely available software packages such as Lisrel (Joreskog and Sorbom, 1993), Mplus (Muthen and Muthen, 1998–2002), Mx (Neale *et al*., 2003), EQS (Bentler, 1995), Genehunter (Kruglyak *et al*., 1996) and Merlin (Abecasis *et al*., 2002). These packages often allow the user to employ path analysis, a technique introduced by Wright (1920, 1921) to formalize relationships between observed traits (phenotypes) and latent (or unobserved) traits (e.g., genotypes). An important feature in the context of genetic analyses of any of the available software packages is the possibility of multi-group analyses. When determining heritability of a trait, the data is usually organized into at least two groups of subjects of different relatedness, e.g., MZ and DZ twins, or adopted and biological siblings. The formal model describing the variance-covariance structure in terms of genetic and environmental (co-) variance is allowed to differ between groups (see e.g., Boomsma and Molenaar, 1986; Eaves and Gale, 1974; Martin and Eaves, 1977).

In addition, multi-group structure is important as heritability is a characteristic of the (sub-) population and may differ across countries, across sex or age cohorts (e.g., Heath *et al*., 1985). Across countries or (sub) populations there may be different allele frequencies (Cavalli-Sforza *et al*., 1994), different genotypic values, or different environmental contributions to the variance. Across sexes, ages or ethnicity there may be different loci that influence the same trait. If heritability estimates are conditional on the value of a quantitative trait such as age or socioeconomic status, the estimates can be obtained across the total scale of a quantitative trait (Purcell, 2002). Ignoring heterogeneity of variance components may lead to biased estimates of these components.

Estimation of quantitative moderation either as effect on the means or on the variance of a trait, necessitates the use of raw data, as opposed to variance-covariance matrices, which is now an option in most SEM packages. An additional advantage of using raw data as input is that it allows a more efficient handling of missing data. Missing data may occur in longitudinal designs, but can also arise if the number of participating family members differs across families.

SEM is not only suited for multi-group structures or using raw data, it also facilitates the analysis of data with relatively complex structures, such as time series with autocorrelated error, non-normal data, multivariate data, incomplete data, or phenotypic interactions.

In this paper we outline a web-based library of statistical scripts (*Mx Scripts Library*, http://www.psy.vu.nl/mxbib), which has been set up in the light of the large international GenomEUtwin collaboration, which is aimed at finding genes for body height, body weight, migraine, longevity, stroke, and cardiovascular disease. An archive for statistical scripts assures synchronized data analysis within this project and allows lucid communication in terms of what models have been tested for a target phenotype. All scripts can be used with the software package Mx (Neale *et al*., 2003). Mx is an algebra interpreter and numerical optimizer for SEM. It is suited for the analysis of both continuous and discontinuous data and can handle input from raw data files, variance-covariance matrices, correlation matrices, or contingency tables.

MX SCRIPTS LIBRARY

The core of the Mx scripts library is a tree structure that helps the researcher to locate a script that can be used for a particular dataset and a specific kind of analysis. For example, for an MZ–DZ twin analysis of a continuous trait aimed at determining heritability, one may choose to analyze continuous data, univariate, variance components, twins only, no sex effects on variance components, 2 group ACE/AE/CE/E and download script rawvc1a.mx (see also Appendix I).

The library allows users to carry out genetic analyses based on phenotypes only, or to carry out genetic analyses which combine phenotypic and molecular genetic information (e.g., SNPs, IBD status). Several (combinations of) options are currently available for analysis through the Mx scripts library. Measured traits can be continuous or discontinuous. For each type of input data, Mx automatically uses the appropriate built-in fit function. For continuous data the default fit function is maximum likelihood (for covariance matrixes) or raw maximum likelihood (for raw data). For contingency tables the default fit function is maximum likelihood assuming bivariate

normal liability. Although a continuous trait is usually preferred over a discontinuous trait for reasons of statistical power, some traits can only be measured on a discontinuous scale. Discontinuous traits can be dichotomous (e.g., affected vs. unaffected) or ordinal (e.g., underweight – normal weight – overweight – obesity – severe obesity), yielding summary counts in contingency tables instead of means and variances/covariances. Contingency tables typically contain the number of (twin) pairs (for each zygosity group) for each combination (e.g., concordant affected, concordant unaffected, discordant). Because of the inherent polygenic background of complex traits, these data are often treated by assuming that an underlying quantitative liability exists with one or more thresholds, depicting the categorization of subjects. Although the liability itself cannot be measured, a standard-normal distribution is assumed for the liability. The thresholds (z-values in the standard normal distribution) are chosen in such a way that the area under the standard normal curve between two thresholds (or from minus infinity to the first threshold, and from the last threshold to infinity) reflects the prevalence of that category. As thresholds may be a function of covariates such as age or sex, contingency tables are not always the optimal input format for analysis. Therefore, the Mx scripts library contains scripts for ordinal data that use either contingency tables or raw data as input.

The basic family structure implemented in the Mx scripts library assumes the classical twin design (i.e., twins only). Most scripts however, are also available for data from twins and any number of additional siblings and some scripts can handle parental data as well. Scripts are available for the analysis of univariate traits, and for bivariate and multivariate traits, (which can be longitudinal), where one can choose between a Cholesky decomposition, Factor models, simplex models or growth curves (see also Neale and Cardon (1992) as a basic resource for describing the types of models that can be fitted with Mx).

All models include specifications for both the covariance structure among relatives (random effects) and for mean structure (fixed effects). Example data files are provided for each script, to enable users to become familiar with a script before applying it to their own data. The scripts are organized in a hierarchical manner such that saturated models vs. variance decomposition models can be tested using likelihood-ratio tests. In addition, scripts are available that assume random samples or selected samples (e.g., DeFries-Fulker regression). Apart from scripts that can be used in heritability studies, several scripts for genetic linkage and association are available.

In all models, tests for sex-limitation may be included. For many analyses scripts are provided for the user to carry out simulation or statistical power analyses. Power analyses are based on finding the non-centrality parameter by constraining (a set of) parameters to zero and refitting the model. The non-centrality parameter is directly related to the sample size required to reject the constrained (i.e., false) model with a specified probability (i.e., power) and a significance level α of 0.05, can be calculated by hand (Hewitt and Heath, 1988; Martin *et al*., 1978) but is also conveniently supplied by Mx.

A specific aim of the Mx scripts library is to provide scripts that implement recent advances in linkage and association analysis. For example, we have included scripts for the simultaneous analysis of linkage and association following the method proposed by Fulker *et al*. (1999). This method also allows for separating a spurious association from a genuine association by comparing trait-allele associations within and between families. An extension to this method, allowing multiple alleles, dominance effects, multiple siblings and the in- or exclusion of parental genotypes, was recently described by Posthuma *et al*. (2004) and is implemented in the scripts in the Mx scripts library.

The website does not only contain scripts, but also provides information on how to restructure data files so they can be read by Mx, some common pitfalls when using specific features of Mx (e.g., when using the definition variable option) as well as batch jobs that can be extremely useful in running the same script for different phenotypes of different genotypes, such as in a full genome screen. The Mx scripts library also provides information on how to import IBD statuses obtained from other programs (such as Genehunter or MERLIN) into Mx.

DISCUSSION

The first successful use of this library within the GenomEUtwin project was recently illustrated by a series of papers describing the genetics of migraine, body height, body mass index, coronary heart disease, stroke and longevity using data from all participating international twin cohorts [see Twin Research, volume 6(5)]. The scripts are posted in the Mx scripts library. In addition we have recently included scripts for sibling interaction (Rietveld *et al*., 2003), rater bias models

(Van der Valk *et al*., 2001; Bartels *et al*., 2003) two-locus linkage (Zhu *et al*., 2004), linkage for threshold characters (Vink *et al*., 2004) and scripts for conducting linkage on data from multiple siblings simultaneously. A limitation of using Mx for linkage applications is that information on IBD status needs to be obtained from other programs. Also, the use of a general SEM package as Mx is not optimal for the genetic analysis of data from large, irregular, pedigrees, spanning several generations.

The Mx scripts library has proven to be useful for SEM users and for geneticists. New scripts are added to the library on a regular basis and specific requests for scripts will be answered when possible. Not only do we aim to continue to add new scripts but we will also continue to add other facilities that will facilitate working with SEM software in general and Mx in particular. As more scripts are being added we plan to make the Mx scripts library searchable in the near future, enabling users to find scripts quickly by typing in keywords.

Researchers in the field of Behavior Genetics are invited to contact the first author to submit their own scripts to the website. Authors of scripts are acknowledged in the title of a script with their initials, as well as in the list of contributors on the website. Relevant references of papers in which a particular statistical model is proposed or applied are also listed on the website.

ACKNOWLEDGMENTS

Financial support was provided by the Netherlands Organization for Scientific Research (NWO, 460-04-033), and by the GenomEUtwin project, which is supported by the European Union Contract No. QLG2-CT-2002-01254.

Appendix I. Current list of scripts available in the Mx Scripts Library. This list changes frequently. The most recent list can also be viewed by clicking *expand entire tree* on http://www.psy.vu.nl/mxbib

Statistical Power	
Univariate	
Significance of a correlation:	power1.mx
Are MZ and DZ correlations significantly different from each other?:	power2.mx
Significance of A or C: continuous data:	power3.mx
Significance of A or C: categorical data:	powcat.mx
Bivariate	
Is a genetic correlation statistically different from 1 or from 0?:	power4.mx
Simulation/calculation	
Simplex:	simulsimplex.mx
Continuous Data	
Univariate	
Saturated	
Twins Only	
No sex effects on variance:	rawSAT1.mx
Sex limitation:	rawSAT2.mx
Selected samples, DeFries-Fulker model	
No sex effects on ACE:	DFuni1.mx
Sex effects on ACE:	DFuni2.mx
Variance components (A, C or D, E - models)	
Twins Only	
No sex effects on variance components	
2 group ACE/AE/CE/E:	rawVC1a.mx
2 group ADE/AE//E:	rawVC1b.mx
Sex limitation	
4 group ACE/AE/CE/E:	rawVC2a.mx
4 group ACE STANDARDIZED:	rawVC2c.mx
4 group ADE/AE/E:	rawVC2b.mx
(5 or) 6 group ACE/ AE/CE/E:	rawVC3a.mx
(5 or) 6 group ACE + different shared env factors across sexes:	rawVC3c.mx
(5 or) 6 group ADE/AE/E:	rawVC3b.mx
Twins and additional siblings	
No sex effects on variance components	
ACE/AE/CE/E models:	rawVC5a.mx
ADE/AE/E models:	rawVC5b.mx

Appendix I *(Continued)*

Variance components and GxE interaction:	rawVCmod1.mx
Variance components and Multiple Raters	
Rater Bias Model:	Raterbias.mx
Psychometric Model:	Psychometric.mx
Variance components and sibling interaction/rater contrast:	Contrast.mx
VC, testing homogeneity of parameters across countries:	rawVC8a.mx
Variance components, testing linkage (A, C or D, E, Q - models)	
Twins only	
pi-hat approach ACEQ/ACE model:	rawVCQ1.mx
IBD mixture distribution approach ACEQ/ACE model:	rawVCQ2.mx
Variance components, multilocus linkage:	multiloc.mx
Variance components, testing linkage, using Mx in BATCH modus	
Generic script:	rawVCQ1b.mx
Download Batch Program:	example_bat.mx
Variance components including linkage and association	
Twins only	
Parental genotypes unavailable	
Testing association:	rawVCla1.mx
Simultaneous Linkage (pi-hat) and Association:	rawVCla2.mx
Simultaneous Linkage (mixture) and Association:	rawVCla3.mx
Parental genotypes available	
Association only:	rawVCla4.mx
Simultaneous Linkage (pi-hat) and Association:	rawVCla5.mx
Simultaneous Linkage (mixture) and Association:	rawVCla6.mx
Bivariate	
Cholesky	
Twins Only	
Two group ACE (rg rc re):	rawVC4a.mx
Two group ADE (rg rd re):	rawVC4b.mx
Twins and additional siblings	
Two group ACE (rg rc re):	rawVC6a.mx
Two group ADE (rg rd re):	rawVC6b.mx
Direction of Causation	
Twins Only:	DOC1.mx
Twins and additional siblings:	DOC2.mx
Multivariate	
Saturated	
Longitudinal/Linear growth model	
Base0:	Base0.mx
Base1:	Base1.mx
Base2:	Base2.mx
Variance components	
Cholesky ACE:	rawVC7b.mx
Multivariate Linkage:	multilink1.mx
Common Pathway:	compath.mx
Independent Pathway:	indpath.mx
Longitudinal/Linear growth model:	Lingrow.mx
Ordinal data	
Univariate	
Saturated	
Contingency tables	
Twins only	
2 group:	ctSATut2.mx
4 group, sex limitation:	ctSATut4.mx
6 group, sex limitation:	ctSATut6.mx
Raw data	
Twins only	
2 group:	ordSATut2.mx
4 group, sex limitation:	ordSATut4.mx
6 group, sex limitation:	ordSATut6.mx

Appendix I *(Continued)*

Variance components	
Contingency tables	
Twins only	
No sex effects	
2 groups ACE:	ctVCut2c.mx
2 groups ADE:	ctVCut2d.mx
Sex limitation	
4 groups ACE:	ctVCut4c.mx
4 groups ADE:	ctVCut4d.mx
5/6 groups ACE:	ctVCut6c.mx
5/6 groups ADE:	ctVCut6d.mx
Raw data	
Twins only	
No sex effects	
2 groups ACE:	ordVCut2c.mx
2 groups ADE:	ordVCut2d.mx
Sex limitation	
4 groups ACE:	ordVCut4c.mx
4 groups ADE:	ordVCut4d.mx
5/6 groups ACE:	ordVCut6c.mx
5/6 groups ADE:	ordVCut6d.mx
Multivariate	
Saturated	
Raw Data	
Twins Only:	ordSATmt2.mx
Variance components	
Raw data	
Twins Only:	ordvcmt2.mx

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**:97–101.

Bartels, M., Hudziak, J. J., Van den Oord, E. J. C. G., Van Beijsterveldt, C. E. M., Rietveld, M. J. H., and Boomsma, D. I. (2003). Co-occurence of aggressive behavior and rule-breaking behavior at age 12: multi-rater analyses. *Behav. Genet.* **33**(5), 607–621.

Bentler, P. M. (1995). *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Boomsma, D. I., Busjahn, A., and Peltonen, L. (2002). The classical twin study and beyond. *Nat. Genet. Rev.* **3**:872–882.

Boomsma, D. I., and Molenaar, P. C. M. (1986). Using LISREL to analyze genetic and environmental covariance structure. *Behav. Genet.* **16**:237–250.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, New Jersey, USA: Princeton University Press.

Eaves, L. J. (1969). The genetic analysis of continuous variation: a comparison of experimental designs applicable to human data. *Brit. J. Math. Stat. Psy.* **22**:131–147.

Eaves, L. J. (1976). A model for sibling effects in man. *Heredity* **36**:205–214.

Eaves, L. J., and Gale, J. S. (1974). A method for analyzing the genetic basis of covariation. *Behav. Genet.* **4**:253–267.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics*. London: Longman.

Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**:259–267.

Heath, A. C., Berg, K., Eaves, L. J., Solaas, M. H., Corey, L. A., Sundet, J., Magnus, P., and Nance, W. E. (1985). Education policy and the heritability of educational attainment. *Nature* **314**(6013), 734–736.

Hewitt, J. K., and Heath, A. C. (1988). A note on computing the chisquare noncentrality parameter for power analyses. *Behav. Genet.* **18**:105–108.

Jinks, J. L., and Fulker, D. W. (1970). A comparison of the biometrical-genetical, MAVA and classical approaches to the analysis of human behavior. *Psychol. Bull.* **73**:311–349.

Joreskog, K. G., and Sorbom, D. (1993). *LISREL 8 User's Reference Guide*. Chicago: Scientific Software International.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**:1347–1363.

Martin, N. G., and Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**:79–95.

Martin, N. G., Eaves, L. J., Kearsey, M. J., and Davies, P. (1978). The power of the classical twin study. *Heredity* **40**:97–116.

Muthén, L., and Muthén, B. (1998–2002). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (2003). *Mx: Statistical Modeling*. 6th edition, Box 980126 MCV, Richmond, VA 23298.

Neale, M. C., and Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. The Netherlands Publishers: Kluwer Academic.

Posthuma, D., de Geus, E. J. C., Boomsma, D. I., and Neale, M. C. (2004). Combined linkage and association tests in Mx. *Behav. Genet.* **34**(2), 179–195.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Res.* **5**(6), 554–71.

Rietveld, M. J. H., Hudziak, J. J., Bartels, M., van Beijsterveldt, C. E. M., and Boomsma, D. I. (2003). Heritability of Attention Problems in Children: I. Cross-sectional Results from a Study of Twins, age 3 to 12. *Am. J. Med. Gen. Part B. (Neuropsychiatric Genetics)* **117B**:102–113.

Van der Valk, J. C., van den Oord, E. J. C. G., Verhulst, F. C., and Boomsma, D. I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: Different views or rater bias?. *J. Child Psychol. Psyc.* **42**(7), 921–931.

Vink, J. M., Beem, A. L., Posthuma, D., Neale, M. C., Willemsen, G., Kendler, K. S., Slagboom, P. E., and Boomsma, D. I. (2004). Linkage analysis of smoking initiation and quantity in Dutch sibling pairs. *Pharmacogenomics J.* **4**:274–282.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *PNAS.* **6**:320–322.

Wright, S. (1921). Correlation and causation. Part 1: Method of path coefficients. *J. Agr. Res.* **20**:557–585.

Zhu, G., Evans, D. M., Duffy, D. L., Montgomery, G. W., Medland, S. E., Gillespie, N. A., Ewen, K. R., Jewell, M., Liew, Y. W., Hayward, N. K., Sturm, R. A., Trent, J. M., and Martin, N. G. (2004). A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. *Twin Res.* **7**(2), 197–210.

Edited by John Hewitt